The incomplete equivalent of paper-and-pencil and computerized versions of the General Aptitude Test Battery

van de Vijver, F.J.R.; Harsveld, M.

Published in:
Journal of Applied Psychology

Publication date:
1994

Journal of Applied Psychology
1994, Vol. 79, No. 6, 852-859

0021-9010/94/$3.00

The Incomplete Equivalence of the Paper-and-Pencil and Computerized Versions of the General Aptitude Test Battery

Fons J. R. Van de Vijver and Menno Harsveld

The performance of 163 applicants for the Dutch Royal Military Academy on the computerized version of the General Aptitude Test Battery (GATB) was compared with the performance of 163 matched applicants on the paper-and-pencil version. There was a modest but clearly discernible influence of computerization. A LISREL analysis showed a reasonable fit for a model postulating two factors that were equally patterned for both test versions. A model postulating equal factor loadings had to be rejected. Individual differences in both the computerized and conventional GATB were strongly related to intelligence. The computerized subtests produced faster and more inaccurate responses than the conventional subtests. Both in terms of number of solved items and correlations with other cognitive measures, the cognitively simple, clerical tests were more affected by computerization than the more complex tasks.

Computerized testing has become increasingly popular during the last decade. If an existing paper-and-pencil test is computerized, the question arises as to whether computerized and paper-and-pencil administration procedures are equivalent. Two kinds of equivalence can be envisaged: qualitative and quantitative (cf. Van de Vijver & Poortinga, 1991). Two instruments show qualitative (or structural) equivalence if they measure the same psychological construct. Qualitative equivalence can be examined by linear structural models or by a factor analysis of the item (or subtest) correlation matrix followed by a target rotation and the computation of Tucker's coefficient of factorial incongruence (e.g., Zegers & Ten Berge, 1985) to assess the degree of factorial similarity.

Qualitative equivalence does not yet imply score comparability. An example is the measurement of temperature in degrees Fahrenheit and Celsius; the two scales measure the same construct but in different measurement units. Numerical score comparability is a condition for quantitative equivalence. Score distributions obtained with quantitatively equivalent instruments should be identical or can be made identical by score transformations such as test equating (e.g., Holland & Rubin, 1982). The definition of *equivalence* offered by the American Psychological Association (1986, quoted in Green, 1991) clearly refers to quantitative equivalence:

> Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode. (Green, 1991, p. 248)

The equivalence of computerized and conventional ability tests has been examined recently in a meta-analysis by Mead and Drasgow (1993). The authors reported a disattenuated cross-mode correlation of .97 for timed power tests (i.e., power tests and tests with a very liberal time limit) and a correlation of .72 for speed tests, pointing to a small or even negligible effect of computerization on timed power tests and a sizeable effect on speed tests. Moreover, computerized tests were found to be slightly more difficult than conventional tests; the difference was 0.03 standard deviation for timed power tests and 0.07 standard deviation for speeded tests (the latter value is not given by the authors, but can be derived from the manuscript).

The reliabilities of computerized versions are almost never lower than those of paper-and-pencil tests (cf. Divgi, 1989; Greaud & Green, 1986).

How should the effects on speeded tests be explained? On the one hand, the difference in scores across test versions could be caused by psychologically peripheral aspects of the stimulus presentation or the response registration procedure that do not threaten the equivalence of the test modes. Mead and Drasgow (1993) attribute the effect on speeded tests to differential motor skills that are required in conventional as compared with computerized testing. On the other hand, validity threatening antecedents of the score differences could also be envisaged. Examples would be an increased test anxiety induced by a computerized administration, a lack of "computer test-wiseness" and of previous exposure to computers or computer-assisted testing, a differential quality of the graphical presentations in the two modes, a presence of multiple items on a single page in the paper-and-pencil version as opposed to one item on the screen at a time in computerized versions, greater ease in looking back at previously solved items in conventional tests, and strategy shifts (e.g., quicker responding in the computer-assisted mode).

Mead and Drasgow (1993) argue that in order to understand the psychological consequences of computerization, confirma-

Fons J. R. Van de Vijver, Department of Social Sciences, Tilburg University, The Netherlands; Menno Harsveld, Royal Netherlands Air Force, The Hague, The Netherlands.

Correspondence concerning this article should be addressed to Fons J. R. Van de Vijver, Department of Social Sciences, Tilburg University, P. O. Box 90153, 5000 LE Tilburg, The Netherlands. Electronic mail may be sent via Internet to fons.vandevijver@kub.nl.

tory factor analytic studies should be carried out, as these studies enable tests of various hypotheses of the impact of computerization. In addition, by studying correlations of conventional and computer-assisted tests with external criteria, their construct validity and, if present, the nature of the test differences can be examined. However, both confirmatory factor analysis and the use of external criteria have been infrequently reported in the literature. Both aspects are taken as a starting point in the present study. We will examine the equivalence of a paper-and-pencil and a computer-assisted version of a speeded aptitude test, namely the General Aptitude Test Battery. The psychological meaning of score differences across the test modes, if present, will be investigated by confirmatory factor analysis and by correlations with other psychological instruments.

Method

Subjects

Subjects were 326 applicants (250 male, 76 female) of the Royal Military Academy in the Netherlands. Their age ranged from 16 to 31 years. All subjects had completed secondary school.

Materials

The General Aptitude Test Battery (GATB) is a general intelligence speed test in multiple choice format, developed by the U.S. Department of Labor. The test contains seven subtests:

1. Name comparison. Two columns of names are presented to the subject who has to indicate whether the names are spelled in a similar or different way. There are 150 items; the time limit is 6 min. The same limit applies to the other subtests unless stated otherwise.

2. Computation. The subtest consists of 50 arithmetic exercises in which whole numbers have to be added, subtracted, multiplied, or divided.

3. Three-dimensional space. Each item consists of a target figure and four drawings of three-dimensional objects. The target figure is depicted as a piece of metal that has to be folded, rolled, or both. Lines in the target figure indicate where the folding should occur. The subject has to indicate which of four drawings of three-dimensional objects can be made out of the target figure. There are 40 items.

4. Vocabulary. Each of the 60 items consists of four words. The subject should indicate which two words have the same or the opposite meaning. The response formats of the conventional and computerized version were different. Whereas in the former the subject should mark two letters on the answer sheet, in the latter all six pairwise comparisons of the four words (first and second word, first and third word, etc.) were presented on the screen.

5. Tool matching. An item is composed of a target figure and four black-and-white drawings of simple tools. The subject should indicate which of the four is the same as the target. The response alternatives all have the same form as the target but differ in their distribution of black and white parts. The subtest contains 49 items, to be completed in at most 5 min.

6. Arithmetic reasoning. The subtest consists of 25 verbally expressed arithmetic exercises. The time limit is 7 min.

7. Form matching. Two sets of line drawings of different shape are presented. For each figure in the first group, the subject is asked to indicate which figure in the second group has exactly the same size and shape. There are 60 figures.

The Dutch version of the test was administered. The second, third, fifth, and seventh subtests of the Dutch version of the GATB are literal translations of the American version; the original stimuli are retained. The other subtests contain minor adaptations to the Dutch language and culture such as the introduction of Dutch names in Name Comparison and money in Arithmetic Reasoning. Both the conventional and computerized Dutch GATB are commercially available instruments that are published by the *Stichting GATB Research Nederland.*

The computerized version of the GATB was administered on an IBM personal computer (Personal System/2, Model 30), including a monitor with EGA color screen and an AT keyboard. The operating system was DOS 4.00. A mold was placed on the keyboard, containing openings only for the arrow keys and the insert, home, page-up, page-down, delete, and end keys. Numerical keys were not used.

Each item was presented separately for the first six subtests, and all figures were presented simultaneously in Form Matching. The correct answer had to be marked by pressing cursor keys. After completion of an item, the subject pressed the page-down key to continue with the next question. The page-up key could be used to scroll back to previous questions and correct answers if desired. Technical details of the computerization of the figure tests can be found in Maarse and Van de Veerdonk (1991).

In addition to the GATB, four other cognitive measures were administered to a subsample of those tested (56 subjects in the computerized mode and 59 subjects in the conventional mode). The tests were administered in a conventional manner. The Berenschot Intelligence Test, consisting of 40 multiple choice items, is a speed test with a time limit of 40 min. In each exercise the subject has to discover the relationship between the verbal, arithmetical, or spatial elements. The Figure Series test was also administered. This is an abstract reasoning test of 50 items; the time limit is 25 min. Series of four figures have to be completed by choosing one out of five possible alternatives. In Instrument Interpretation, a paper-and-pencil test of spatial skill, the candidate has to select the correct aircraft position given information on the horizon and compass out of a set of five alternatives. There is a time limit of 20 min to solve the 60 items. The fourth instrument was the Determinationsgerät, a test of reaction speed. Two kinds of stimuli are used, namely visual and auditory. The subject should react as fast as possible to the presentation of the stimulus by pressing a button by hand or a pedal by foot, depending on the kind of stimulus. The test duration is 150 s and the score is the number of correct responses.

Finally, a questionnaire to evaluate the computerized procedure was administered, containing questions about the clearness of instruction, readability of the screen, use of back-scrolling to correct answers, and experience with computers.

Procedure

The computerized version of the GATB was administered to half of the sample, the paper-and-pencil version to the other half. The two groups were matched for age (computer group: $M = 18.3$ years, $SD = 0.96$; paper-and-pencil group: $M = 18.5$, $SD = 1.00$; $p > .05$), sex (the proportion of males was .77 in both groups; $p > .05$), and general intelligence (Berenschot intelligence score of the computer group: $M = 31.69$, $SD = 4.59$; paper-and-pencil group: $M = 31.72$, $SD = 4.59$; $p > .05$).

Results and Discussion

Number of Items Solved and Skipped

The number of items solved (either correctly or incorrectly), the proportion of correctly solved items, and the number of skipped items were determined for each subtest. It can be seen in Table 1 that the two versions revealed different numbers of

Table 1
Average Number of Solved Items, of Correctly Solved Items, and of Skipped Items per Subtest and Test Version

Subtest and test mode	Items answered	Items correctly solved	Proportion correctly solved of items answered	Items skipped
Name Comparison				
P & P				
M	78.3*	75.5*	.97*	.07
SD	14.9	14.3	.04	.65
Comp				
M	91.9	87.0	.95	.00
SD	19.0	16.7	.03	.00
Computation				
P & P				
M	26.7*	24.3	.91*	.83*
SD	3.7	3.7	.08	2.33
Comp				
M	27.8	24.6	.89	.00
SD	3.5	3.8	.09	.00
Three-Dimensional Space				
P & P				
M	29.4*	26.2*	.89*	.68*
SD	4.5	4.6	.09	1.77
Comp				
M	31.1	24.4	.78	.00
SD	5.2	5.3	.11	.00
Vocabulary				
P & P				
M	38.4*	32.1*	.84*	1.22*
SD	5.4	5.0	.08	1.84
Comp				
M	37.1	30.0	.81	.00
SD	5.8	5.1	.09	.00
Tool Matching				
P & P				
M	38.3*	36.8*	.96	.08*
SD	5.6	5.2	.05	.33
Comp				
M	42.2	40.2	.95	.00
SD	6.0	5.7	.03	.00
Arithmetic Reasoning				
P & P				
M	18.2*	16.6	.91*	.18*
SD	2.5	2.7	.08	.46
Comp				
M	18.9	16.6	.88	.00
SD	2.7	2.8	.09	.00
Form Matching				
P & P				
M	37.6*	35.4*	.94*	.72*
SD	7.0	7.0	.06	2.29
Comp				
M	36.2	31.4	.87	.00
SD	6.0	5.5	.08	.00

Note. P & P = paper-and-pencil version; Comp = computerized version.
* $p < .05$ between the paper-and-pencil and the computerized subtest (i.e., between the average with the asterisk and the average immediately below).

solved items for each subtest. The computerized version showed higher average numbers of solved items for Name Comparison, Computation, Three-Dimensional Space, Tool Matching, and Arithmetic Reasoning; higher scores on the conventional subtests were obtained for Vocabulary and Form Matching. The slower responding on Vocabulary is probably caused by the difference in response format (see Method). The presentation of all word pairs in the computerized version might have increased the time needed to read the item or made an exhaustive search (i.e., a comparison of all pairs) rather than a quicker, self-termi-

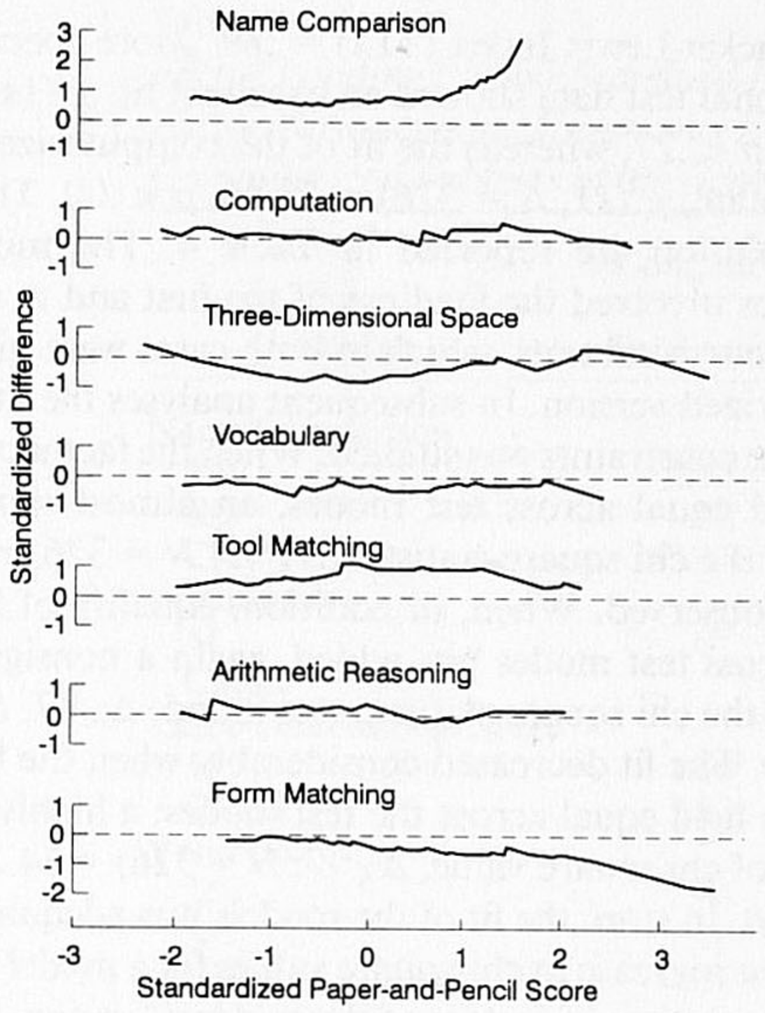

Figure 1. Equipercentile equating curves of the computerized and conventional tests. (Standardized conventional test scores are given on the horizontal axis; the vertical axis represents the difference of the computerized and conventional test divided by the standard deviation of the conventional test.)

nating search (i.e., a comparison of pairs until a suitable answer is found) more likely.

The increased speed of responding in the computerized mode might be due to the quicker administration (stimulus presentation and response registration) by the computer as compared with the paper-and-pencil version (e.g., no pages have to be turned over in the computerized version). However, there were indications that the increased speed of responding was also generated by differences in solution strategies. All subtests produced higher proportions of correctly solved items in the conventional version than in the computerized version (see Table 1). This finding would not have been expected if speed of administration was the only reason of the differential performance in the two modes. The two test versions induced different levels of speed and accuracy; the computerized version induced quicker and the paper-and-pencil version more accurate responding. Faster and more inaccurate responding to a computerized test version has also been reported by Neubauer, Urban, and Malle (1991).

The number of skipped items was consistently different in both versions; the paper-and-pencil version invariably showed a larger proportion of skipped items (Table 1). No subjects skipped items in the computerized version, even though the software did not prohibit this. The difference in number of skipped items reached significance for all subtests except for Name Comparison (all $ps < .05$). The psychological background of the difference is equivocal. Because the computer controlled the item presentation, item skipping required a deliberate choice by the examinee; unintended item skipping (e.g., when the subject first solves the second item that is printed on a page and then forgets to solve the first item) was much more likely in the paper-and-pencil version than in the computerized version. In addition, it could well be that subjects did not skip items in the computerized version because they were not instructed that they could do so. It was not mentioned during the instruction of the paper-and-pencil tests either but the opportunity was probably more obvious.

It could be speculated that the quicker though less accurate responding in the computer-assisted condition was a result of a perceived demand characteristic of the testing situation. Computers are often associated with high-speed performance. This impression may have been reinforced by the speed of performance of the computer in the testing situation (e.g., drawing a display, responses to cursor movements and scrolling commands). It could well be that individuals adapt their behavior accordingly. This demand set may be related to computer wiseness. The differences of the two test versions will probably diminish when examinees have acquired more experience in computerized testing.

The consequences of this strategy difference on the total number of correctly solved items in the two conditions were varied. Two subtests revealed higher averages in the computerized version (viz. Name Comparison and Tool Matching), two subtests showed equal results (viz. Computation and Arithmetic Reasoning), and three means were higher in the conventional version (viz. Vocabulary, Three-Dimensional Space, and Form Matching). These findings can also be expressed in terms of

Table 2
Correlations Between the Subtests (Computerized Version Above the Diagonal; Paper-and-Pencil Version Below the Diagonal)

Subtest	1	2	3	4	5	6	7
1. Name Comparison	—	.37*	.25*	.50*	.45*	.26*	.31*
2. Computation	.15	—	.07	.39*	.26*	.63*	.00
3. Three-Dimensional Space	.03	.10	—	.46*	.34*	.14	.44*
4. Vocabulary	.10	.13	.24*	—	.38*	.43*	.31*
5. Tool Matching	.22*	.24*	.35*	.22*	—	.11	.45*
6. Arithmetic Reasoning	.14	.55*	.22*	.31*	.24*	—	.08
7. Form Matching	.25*	.22*	.41*	.36*	.50*	.23*	—

* $p < .05$.

effect sizes (i.e., the difference of the averages of the computerized and conventional tests divided by the pooled within-group standard deviation). The average effect size of all seven tests was 0.01, and the standard deviation was 0.48. So, the overall impact was low though highly dissimilar for the subtests.

The present study shows that a data analysis of a speed test in an equivalence study that considers merely the number of correct answers is incomplete. The proportion of solved items, of skipped items (if applicable), and of correctly solved items should also be studied in order to detect strategy differences across the administration modes.

Equipercentile equating curves have been presented in Figure 1. The largest deviations from a simple, linear transformation rule were found for the first and last subtests, caused by a few very high scoring subjects in one mode, not present in the other mode; for example, a few subjects obtained very high scores on the computerized version of Name Comparison but such extreme scores were not present in the conventional mode.

Covariance Modeling

The equivalence of the computerized and conventional subtests was investigated by LISREL; subtest correlations are presented in Table 2. The hypothesis of equal covariance matrices in both test modes had to be rejected, $\chi^2(28, N = 326) = 69.51$, $p < .00$ (see Table 3). The lack of fit was further explored by fitting increasingly constrained models to the data; a similar procedure has been described by Vandenberg and Self (1993). In the first analysis a test of the same number of factors on both versions was examined. The number of factors and the patterning of free and fixed zero loadings were derived from pilot analyses of data of applicants not considered here and of data reported in the GATB manual (United States Department of Labor, 1970, p. 31). The first latent variable is related to perceptual speed; the second latent variable, arithmetic ability, was mainly defined by the arithmetic subtests. In addition to the zero loadings, two loadings with a value of one were fixed in order to circumvent problems of identifiability, namely the loading of Three-Dimensional Space on the first factor and Computation on the second factor.

The model fit the data fairly well, $\chi^2(22, N = 326) = 49.84$, $p < .00$; Root Mean Square Error of Approximation (RMSEA) = .06; Tucker-Lewis Index (TLI) = .89. More specifically, the conventional test data showed an excellent fit, $\chi^2(11, N = 326) = 13.37$, $p < .27$, whereas the fit of the computerized tests was less adequate, $\chi^2(11, N = 326) = 36.46$, $p < .00$. The loadings of the solution are reported in Table 4. The most striking differences involved the loadings of the first and to a lesser extent the fourth subtests, which in both cases were higher in the computerized version. In subsequent analyses the fit of models with more constraints was studied. When the factor correlations were held equal across test modes, an almost significant increase of the chi square statistic, $\Delta\chi^2(3, N = 326) = 7.08$, $p < .07$, was observed. When, in addition, equality of factor variances across test modes was added, again a nonsignificant increase of the chi square statistic was found, $\Delta\chi^2(7, N = 326) = 10.69$, *ns*. The fit decreased considerably when the factor loadings were held equal across the test modes; a highly significant increase of chi square value, $\Delta\chi^2(7, N = 326) = 34.28$, $p < .00$, was found. In sum, the fit of the models was adequate as evaluated by the increase in chi-square values for a model postulating the same number of factors of the two test versions; however, a model of equal factor loadings had to be rejected.

An additional analysis of the psychological similarities and differences of the two test versions was obtained by correlating the scores on the (conventional and computerized) subtests with the scores on the Berenschot Intelligence Test, Figure Series, Instrument Interpretation, and Determinationsgerät. The correlations are presented in Table 5. Of the four cognitive measures, the Berenschot Intelligence Test showed the strongest correlations with both the computerized and conventional subtests. This pattern suggests that individual differences in both the computerized and conventional GATB are strongly related to intelligence.

The correlations of the four measures with Computation, Three-Dimensional Space, Vocabulary (despite the difference of presentation format of both the test versions), and Tool Matching did not differ markedly for the conventional and computerized subtests. The most salient differences in correlations were found for Name Comparison and Form Matching and, to a lesser degree, for Computation. Name Comparison showed a stronger relationship with intelligence in the computerized version than in the conventional version ($p < .05$; see Table 5); the relationship of Form Matching subtest scores with reaction

Table 3
Results of the LISREL Analysis

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	TLI	RMSEA
Equal covariance matrices	69.51*	28	NA	NA	NA	.07
Null model[a]	547.92*	42	NA	NA	NA	.19
Nested models						
Equal factor model	49.84*	22	NA	NA	.89	.06
Equal factor covariances	56.92*	25	7.08	3	.89	.06
Equal factor variances	67.61*	32	10.69	7	.91	.06
Equal factor loadings	101.89*	39	34.28*	7	.87	.07

Note. TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; NA = not applicable.
[a]Model specification of null model for both test versions (in LISREL notation): LX = identity matrix; TD = zero matrix; phi has free diagonal elements and zero off-diagonal elements.
* $p < .05$.

Table 4
Factor Loadings, Their Standard Errors, Common Metric Standardized Solution (CMSS), Factor Variances and Covariances, and Uniquenesses of the Paper-and-Pencil and Computerized Subtests of the General Aptitude Test Battery

	Computer-assisted version		Paper-and-pencil version		Uniqueness	
Subtest	Perc. speed	Arithm. abil.	Perc. speed	Arithm. abil.	Comp	P & P
Name Comparison						
M	2.53	1.68	1.54	.47	165.84	186.60
SE	.52	.46	.64	.60		
CMSS	7.15	4.54	4.34	1.25		
Computation						
M	.0[a]	1.00[a]	.0[a]	1.00[a]	4.92	8.42
CMSS		2.70		2.70		
Three-Dimensional Space						
M	1.00[a]	.0[a]	1.00[a]	.0[a]	16.99	15.35
CMSS	2.83		2.83			
Vocabulary						
M	.84	.66	.74	.44	11.80	20.01
SE	.16	.14	.23	.21		
CMSS	2.36	1.79	2.09	1.18		
Tool Matching						
M	1.14	.0[a]	1.40	.0[a]	18.85	16.23
SE	.20		.27			
CMSS	3.21		3.97			
Arithmetic Reasoning						
M	.0[a]	.70	.0[a]	1.10	3.20	1.37
SE		.10		.30		
CMSS		1.88		2.97		
Form Matching						
M	1.09	.0[a]	2.36	.0[a]	18.55	17.73
SE	.19		.45			
CMSS	3.09		6.67			
Factor variances and covariances						
Perceptual speed						
M	10.38		5.60			
SE	2.80		1.84			
CMSS	1.30		.70			
Arithmetic ability						
M	2.15	9.58	1.98	4.97		
SE	1.13	1.96	.79	1.71		
CMSS	.28	1.32	.26	.68		

Note. Perc. speed = perceptual speed; Arithm. abil. = arithmetic ability; Comp = computerized version; P & P = paper-and-pencil version.
[a] Fixed loading.

speed was significantly higher in the paper-and-pencil condition ($p < .05$). It can be tentatively concluded that the cognitively simple clerical tests were more affected by computerization than the more complex tasks (cf. Greaud & Green, 1986; Henly, Klebe, McBride, & Cudeck, 1989).

The slight change of nature of the GATB factors after computerization might be generated by the differential acquaintance with both stimulus modes, the same factor that presumably generated the differences in strategy observed earlier. Individuals will be more familiar with the conventional than with the computerized test version. There is a rich literature showing that the nature of a psychomotor task can vary as a function of previous experience (e.g., Ackerman, 1987; Fleishman & Hempel, 1954). In a classical series of experiments, Fleishman and Hempel (1954) found that intellectual differences on a psychomotor task had to be accounted for by intellectual factors in the beginning of the training, but after prolonged training motor abilities became prominent. Analogously, our subjects were probably novices in dealing with computerized tests and they had more experience with conventional tests. In this interpretation, computer test-wiseness shows considerably more individual differences than paper-and-pencil test-wiseness; repeated administration of the computerized tests should give rise to a reduction of the differences in test scores and in correlations with external measures.

Questionnaire

Almost all subjects found the readability of the screen adequate. Fifteen percent said that the screen images were not

Table 5
Correlations Between the Computer-Assisted (N = 56) and Conventional (N = 59) General Aptitude Test Battery Subtest Scores and Additional Cognitive Measures

	Berenschot Intelligence		Figure Series		Instrument Interpretation		Determinationsgerät	
Subtest	Comp.	P & P	Comp.	P & P	Comp.	P & P	Comp.	P & P
Name Comparison	**.52***	**−.16**	.11	−.10	**.25**	**−.14**	.17	.17
Computation	.50*	.29*	−.01	.01	**.41***	**−.05**	.29*	.16
Three-Dimensional Space	.32*	.38*	.01	.04	.04	.16	−.02	.06
Vocabulary	.52*	.47*	.12	−.09	.14	.25	.23	.28*
Tool Matching	.33*	.13	.11	.13	.07	.34*	.23	.37*
Arithmetic Reasoning	.38*	.51*	.02	.04	.46*	.20*	.21	.26*
Form Matching	.13	.16	.02	.19	**−.16**	**.24**	**.09**	**.50***

Note. Comp = computerized version; P & P = paper-and-pencil version. Boldface correlations differ significantly from each other ($p < .05$, two-tailed).
* $p < .05$.

clear; the most common complaints involved the deformation of figures and the occurrence of jagged edges in drawings of straight lines. No subjects reported problems with the clarity of the test instructions or the operation of the keyboard. Two-thirds of the subjects had used the scroll-back and correction facilities.

Only 9% did not like to work with a computer. Three quarters of the subjects had computer experience of any kind. Their subtest scores did not differ significantly from those of subjects without computer experience.

Conclusion

The equivalence of a computerized and a paper-and-pencil version of the GATB was investigated. Overall, the differences between the test versions were small though noticeable. A LISREL analysis showed a reasonable fit for a model postulating two factors that were equally patterned for both test versions. Individual differences in both the computerized and conventional GATB were strongly related to intelligence. However, there were also differences between the computerized and conventional tests. The computerized subtests produced faster and more inaccurate responses than the conventional subtests. The cognitively simple clerical tests were more affected by computerization than the more complex tasks.

The present findings have implications for equivalence studies. First, equivalence of paper-and-pencil and computerized versions of a test should be demonstrated rather than assumed (Van de Vijver, 1987). Based on previous (e.g., Mead & Drasgow, 1993) and current results, there is no reason to be optimistic about the equivalence of conventional and computerized speed tests, particularly in the case of simple tests.

Second, possible changes in the nature of the stimulus material and in the solution strategies adopted by those tested should be acknowledged in the design and analysis of equivalence studies. This can be achieved by a detailed analysis of the response patterns and of the nomological network of both test versions. Computerization can induce shifts in the nature of the tasks and the strategies used by those taking the test that can easily remain undetected. Techniques to detect differential item functioning (e.g., Holland & Thayer, 1988) and test equating (e.g., Wainer, 1990) can enable a fine-grain analysis of the differences.

Finally, a lack of equivalence does not imply that computer-assisted tests would have a lower predictive validity than paper-and-pencil tests. In our view, there is no a priori reason to assume any systematic effect of computerization on the predictive validity of a test.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3–27.

Divgi, D. R. (1989). Estimating reliabilities of computerized adaptive tests. *Applied Psychological Measurement, 13,* 145–149.

Fleishman, E. A., & Hempel, W. E. (1954). Change in factor structure of a complex psychomotor test as a function of practice. *Psychometrika, 19,* 239–252.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10,* 23–34.

Green, B. F. (1991). Guidelines for computer testing. In T. B. Gutkin & J. C. Conoley (Eds.), *The computer and the decision making process. Buros-Nebraska symposium on measurement and testing* (pp. 245–273). Hillsdale, NJ: Erlbaum.

Henly, S. J., Klebe, K. J., McBride, J. R., & Cudeck, R. (1989). Adaptive and conventional versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement, 13,* 363-371.

Holland, P. W., & Rubin, D. B. (Eds.). (1982). *Test equating.* New York: Academic Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Maarse, F. J., & Van de Veerdonk, J. L. A. (1991). Graphical representations in computerized psychological tests. In L. J. M. Mulder, F. J. Maarse, W. P. B. Sjouw, & A. E. Akkerman (Eds.), *Computers in psychology. Applications in education, research and psychodiagnostics* (pp. 62–67). Lisse: Swets & Zeitlinger.

Mead, A. L., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

Neubauer, A. C., Urban, E., & Malle, B. F. (1991). Raven's Advanced Progressive Matrices: Computerunterstützte Präsentation versus Standardvorgabe [Raven's Advanced Progressive Matrices: Computer-assisted presentation versus standard administration]. *Diagnostica, 37,* 204–212.

United States Department of Labor (1970). *General Aptitude Test Battery: Section III. Development.* Washington, DC: U.S. Government Printing Office.

Van de Vijver, F. J. R. (1987). Het gebruik van computer-ondersteunde tests in de diagnostische praktijk [The use of computer-assisted tests in the assessment practice]. *De Psycholoog, 22,* 10–15.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Dordrecht: Kluwer.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitments to the organization during the first 6 months of work. *Journal of Applied Psychology, 78,* 557–568.

Wainer, H. (Ed.) (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Erlbaum.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika, 50,* 17–24.

Received April 30, 1993
Revision received May 3, 1994
Accepted May 9, 1994 ■

P&C Board Appoints Editor for New Journal: *Psychological Methods*

The Publications and Communications Board of the American Psychological Association has appointed an editor for a new journal. In 1996, APA will begin publishing *Psychological Methods*. Mark I. Appelbaum, PhD, has been appointed as editor. Starting January 1, 1995, manuscripts should be directed to

Mark I. Appelbaum, PhD
Editor, *Psychological Methods*
Department of Psychology and Human Development
Box 159 Peabody
Vanderbilt University
Nashville, TN 37203

Psychological Methods will be devoted to the development and dissemination of methods for collecting, understanding, and interpreting psychological data. Its purpose is the dissemination of innovations in research design, measurement, methodology, and statistical analysis to the psychological community; its further purpose is to promote effective communication about related substantive and methodological issues. The audience is diverse and includes those who develop new procedures, those who are responsible for undergraduate and graduate training in design, measurement, and statistics, as well as those who employ those procedures in research. The journal solicits original theoretical, quantitative empirical, and methodological articles; reviews of important methodological issues; tutorials; articles illustrating innovative applications of new procedures to psychological problems; articles on the teaching of quantitative methods; and reviews of statistical software. Submissions should illustrate through concrete example how the procedures described or developed can enhance the quality of psychological research. The journal welcomes submissions that show the relevance to psychology of procedures developed in other fields. Empirical and theoretical articles on specific tests or test construction should have a broad thrust; otherwise, they may be more appropriate for *Psychological Assessment.*